
Douglas R. Hofstadter, Professor of Cognitive Science and Computer Science
Adjunct Professor of Psychology, Philosophy, and History and Philosophy of Science
Center for Research on Concepts and Cognition
Indiana University • 510 North Fess Street
Bloomington, Indiana 47408
(812) 855-6965

September 24, 1990

Dr. Paul Berg
Beckman Institute, Room B062
Stanford University Medical Center
Stanford University 94305

Dr. Maxine Singer
Carnegie Institution
1530 P Street N.W.
Washington D. C. 20005

Dear Paul and Maxine,

I really should know myself better... I should know that when I decide to get involved in reading a book manuscript, I will wind up spending huge amounts of time on it, putting red all over the place, getting involved on every level, from minute typographical things to global organization. In any case, I had a great time, but now — at long last — I'm done with it!

First of all, I want to apologize to you, both for the amount of red (or black, depending on which copy you get) and for the sometimes rather impatient or annoyed tone of my remarks in the margins. I seem to have a very opinionated streak — I like things to be said in certain ways, and I think certain phrases sound wrong, and so on and so forth. I hope you'll take the tone of my remarks with a grain of salt — I have the greatest of admiration for what you've done here, and have enjoyed reading it very much. Needless to say, I've also learned a lot (though I certainly haven't yet absorbed all of it).

As you will see, I am a fanatic about clarity in writing — not only clarity of imagery, but also syntactic non-ambiguity. This means that I often suggest rewordings of passages that may look completely fine to an author who already knows what is intended; however, that same passage may have been hard to parse for me, an outsider, and so I have tried to eliminate that alternate parsing.

A perfect example of this is on page 67, where you write, "In practice, visualizing a pattern of restriction nuclease fragments requires separating the mixture of fragments according to length." First there was "in practice"; was this meant as the opposite of "in theory"? That's what it means to me. However, given the context, I decided that what you probably meant was "in the laboratory". Next came the term "visualizing", which to me has just one meaning — "producing visual imagery" — and it was obvious you didn't mean that. I decided you meant essentially "making [something] visible". Finally, even though I had just read the previous several sentences about how restriction nucleases chop up DNA, I was totally thrown by the phrase "restriction nuclease fragments", until I realized that what you meant was not "fragments *of* restriction nucleases", but "fragments *produced by* restriction nucleases" — and in fact, not by a *bunch* of restriction nucleases, but by a *particular* one. Putting it all together, I came up with the following rewrite: "In the laboratory, in order to actually see the pattern of DNA fragments produced by a particular restriction nuclease, one must separate the various fragments in the mixture according to their

lengths.” It’s a little longer, admittedly, but it sure makes it easier on the lay reader!

Scientists who know what they mean and want to write concisely tend to use lots of “code phrases” whose meanings are completely obvious to a colleague, but which risk evoking completely unwanted meanings in the mind of a non-specialist. For example, on page 75, you use the term “eukaryotic virus”. When someone who is completely self-confident about the meanings of both words first hears it or reads it, they pretty much *have* to come to the conclusion that it refers to *viruses that attack eukaryotic organisms*, and as soon as they’ve thus “decoded” it once, it becomes a totally comfortable, problem-free little phrase that they’ll probably start using on their own. Very nice. However, to a lay person encountering it for the first time, someone whose mastery of the two terms is somewhat shaky, it will be a confusing phrase. They will think, “How can a virus be eukaryotic? It’s not even a cell! How can it have a nucleus? Did I miss something crucial? Are there some *very big viruses* that somehow have nuclei?” And so on.

All these concise little code phrases just add up. “Expression vector”. “Restriction nuclease digest”. “Anonymous probe”. “Retrotransposon”. “Viral oncogene”. “Transducing phage”. “Insertional mutagenesis”. It’s not that I’m opposed to technical terms — not at all. But there are certain terms that somehow seem to me to be completely clear (perhaps only in retrospect, now that I’m used to them), whereas others seem somehow *too concise*, and therefore opaque. To me, they seem to *impede* communication instead of aiding it. This is a subtle matter, and I’m not sure exactly what my recommendation is. Probably there’s no simple recipe. I’ll just let the suggested rewrites speak for themselves — maybe collectively they’ll convey to you my feelings on this matter.

My passionate drive for syntactic non-ambiguity has led me to become a big fan of dashes (em-dashes, that is). When they are used properly, their “feel” is just about halfway between that of commas and that of parentheses, which makes them extremely helpful for purposes of clarity. Very often, an appositive phrase set off by commas causes a good deal of potential ambiguity, whereas with dashes it would be absolutely clear. For example, on page 56 of Chapter 4, it says, “Phage, bacterial viruses, can also bring foreign DNA...” The way it’s written could lead a non-savvy reader to think that you’re talking about phage *and* bacterial viruses, whereas with a dash it’s obvious that it’s an appositive phrase: “Phage — bacterial viruses — can also bring foreign DNA...” Similarly, on page 20 of Chapter 1, you write, “Almost a century after their independent beginnings, **the three separate scientific fields, chromosome behavior and structure, abstract genetic analysis, and biochemistry** were unified.” As it is, with commas, the part in boldface might be read as a list whose first item is “the three separate scientific fields”. With dashes, it is much clearer: “Almost a century after their independent beginnings, the three separate scientific fields — chromosome behavior and structure, abstract genetic analysis, and biochemistry — were unified.”

Note that em-dashes are really very much like parentheses, in that they tend to go in pairs. In fact, in some European books, I have seen them treated almost exactly like parentheses. For example, the previous example would be typeset as follows: “Almost a century after their independent beginnings, the three separate scientific fields —chromosome behavior and structure, abstract genetic analysis, and biochemistry— were unified.” Note how there are spaces on just one side of each dash, telling you whether it is a *left* dash or a *right* dash. While I like this convention, I wouldn’t go so far as to recommend it in your book.

By the way, real em-dashes are not hyphens, although the dashes in your manuscript were all typed as hyphens. When I learned typing, I was taught to use a *pair* of hyphens flanked by blanks, although a more common convention leaves the blanks out. I'm sure your publisher will take care of such details, however.

A related ambiguity-connected passion of mine concerns compound modifiers; I feel they should almost always be hyphenated. For example, on page 86 you start a paragraph with "By the late 1970's, it was clear that the **protein coding sequences** in a eukaryotic gene...". A non-specialist could easily interpret this, at least at first, as "it was clear that the protein, coding sequences in a eukaryotic gene, ..." With a hyphenated modifier, it becomes impossible to misinterpret: "By the late 1970's, it was clear that the **protein-coding sequences** in a eukaryotic gene..." On page 92, another paragraph starts, "For example, the maturation of an **RNA polymerase II primary transcript** of a **protein coding gene** into a messenger RNA requires several steps." Here you *really* could benefit from hyphens: "For example, the maturation of an **RNA-polymerase-II primary transcript** of a **protein-coding gene** into a messenger RNA requires several steps." Even here, though, using a long compound noun ("RNA polymerase II") as a modifier in front of "primary transcript" is quite confusing. Why not spell the whole idea out a little less concisely and a little more directly? It could go something like this: "For example, it requires several steps for the primary transcript of a protein-coding gene, transcribed by RNA polymerase II, to mature into a messenger RNA."

I have the impression, not just from your book, that usage of multiple complicated compound modifiers is rampant in molecular-biology articles. This may not confuse technical colleagues, but it can wreak havoc with less with-it readers. My favorite example of a problem with compound modifiers in your manuscript is found on page 50, where you refer to "**a ribosome-transfer RNA-mediated process**". When I first hit this, I scratched my head and wondered, "*What* ribosome-transfer process? I don't recall any such thing. What do they mean? And although I admit the process they were just talking about could be called 'RNA-mediated', I wouldn't have put it that way..." Then all of a sudden it hit me that you were referring to *transfer RNA*, and that what you meant was essentially "a ribosome-and-tRNA-mediated process". However, a far better way to phrase it would be "**a process mediated by ribosomes and transfer RNA**". The blank space between "transfer" and "RNA" in your original phrase, however, threw me totally off, making me perceive two hyphenated compound modifiers, even though I have read a million times about tRNA, and even lectured and written about it numerous times.

By the way, while we're on the topic of RNA, I myself prefer writing "tRNA", "rRNA", and "mRNA" to "transfer RNA", "ribosomal RNA", and "messenger RNA" — they're more concise and each one of them ought to be an independent concept in the reader's mind. It's sort of like "DNA" and "RNA" — you wouldn't want to write out their full names every time. I'm not suggesting that readers not be told what the "t", "r", and "m" stand for — just that you ought to allow yourselves to use those abbreviations in spots, where it would make things a little easier.

On page 176 you write, "The introduction of a different oncogene, *myc*, under the control of a mouse mammary tumor virus promoter, yields mammary tumors." I found the phrase "a mouse mammary tumor virus promoter" hard to decipher. Here's a suggested rewrite: "Mouse mammary tumor virus promoter controlled alternate oncogene introduction yields mammary tumors." (Just kidding!)

My most outspoken comments, I think, come at the top of page 41, where, on the topic of the genetic code, you wrote the following: "The challenge was to decipher the code and to learn how it is translated from DNA to protein." Firstly, I object to the word "decipher" applied to the genetic code. "Crack the genetic code", okay — but "decipher"? No way! That verb applies to *messages written in a code*, not to the code itself. Thus during World War II, the British deciphered German messages sent to the U-boats, and they did so by cracking the code in which those messages were written. But they didn't *decipher* the code. (I feel like William Safire, I must say!) But a worse sin was that you then went on and said, "to learn how it [the genetic code itself] is translated from DNA into protein". Shame, shame! It reminds me of when people say things like, "They discovered that genes were genetic codes for proteins." I cringe when I hear such things!

So much for syntactic aspects. One of my more important content-related concerns had to do with the way in which, in the first few chapters, you slide into talking about DNA, genes, and chromosomes as "information". It seems to me that historically, there were roughly four distinct stages of the concept of "gene". They go something like this:

- (1) a gene as an abstract entity responsible for a particular inherited trait (Mendel)
- (2) a gene as a physical piece of a particular chromosome (T. H. Morgan)
- (3) a gene as *correlated*, on a holistic level, with a particular protein (but no real sense of *coding* — i.e., one-to-one sequential matchup of pieces of the gene with pieces of the protein) (Garrod, Beadle, Tatum)
- (4) a gene as a linear chain *coding* for a particular protein in a sequential, piece-by-piece manner

My feeling is that you start using words like "information" and "encode" when you have gotten readers to stage 3 but not yet to stage 4. To me, this doesn't work. When all one knows is that particular genes are correlated with particular proteins as *wholes*, one won't think of genes as constituting *information*. There's no sense of "reading" a gene until you think of genes as linear structures composed of letter-like entities, and proteins as other types of linear structures composed of some other kind of letter-like entities. Then and only then, in my opinion, does it start to make sense to talk in terms of "coding" and "information".

There is a related comment on page 40 about the word "express", which I'll retype here just so you can think about it in advance. What I wrote was essentially this: "When a normal person speaks about how a message is expressed, they mean how it is *put into the medium* (here the medium is DNA, but in general, it is spoken or written language). By contrast, when a molecular biologist talks about the expression of a message, they mean how it is *gotten out of the medium* (DNA or RNA)." Thus there is a curious contrast between everyday and technical uses of the same word. The way I see it is that in biology, there are really two media — nucleotides and amino acids — and you are taking an input message in *one* medium and expressing it in the *other* medium. It's just that no one speaks of proteins as "messages"; only genes (and possibly mRNA) are described as "messages".

You'll notice that I object numerous times to the word "particle" to refer to things like viruses, ribosomes, etc. It's probably just my physics background showing through, but I've never liked that word in biological contexts, because to me, "particle"

carries very strong connotations of indivisibility and fundamentality, as epitomized by electrons, quarks, and so on.

Another minor hobbyhorse of mine is the term “catalyze”, in reference to the function of enzymes. Although strictly speaking, it is the right term to use, it conjures up entirely the wrong imagery for me (and, I would suspect, for most lay people) — namely, the idea that a given enzyme’s presence mildly increases the speed of some reaction (doubling or tripling it, say), when in fact the enzyme is so phenomenally catalytic that it speeds the reaction up by a factor of a million or a billion! For this reason, I prefer simply to say that an enzyme *carries out* a reaction, even if strictly speaking it’s not true. Everybody speaks of enzymes as “cellular machinery”, and that’s what they mean!

An overall comment on the diagrams: they are not caricatural enough for my taste. They also contain too many technical terms and symbols for me. What I want to see is a situation stripped down to its bare essence, rendered maximally simple and maximally clear, rather than clothed in lots of details. And, incidentally, I think your captions are generally too terse — I would like to see them essentially self-sufficient, as in *Scientific American* articles.

} like
Bowman

Finally, I come to a few topics that I’d like to see you expand on. My absolute favorite one is *overlapping genes*. Ever since I first read about such things in ϕ X174 DNA, I was fascinated. The whole idea is so weird, so much like biological puns or *double entendres*. There are apparently two types — the shifted-reading-frame type (as in ϕ X174), and the nonsense-strand type (i.e., where the two genes face each other on complementary strands). Both types are so amazing and so virtuosic. How common is this phenomenon? Does it happen only in viruses? Does it ever happen in prokaryotes or eukaryotes? How could such tricks have evolved?

A related question is the extent to which regulatory sequences (including out-of-frame start codons and stop codons) might appear *by accident* inside coding sequences. Why shouldn’t this happen occasionally — in fact, quite often? And how would inducers, repressors, RNA polymerases, and other related enzymes recognize such “accidents” for what they were and ignore them?

A question related to this one is how mistakes in copied DNA are recognized and corrected. How in hell can a dumb little enzyme inspect an isolated piece of DNA and know that it contains an error? Or doesn’t it happen that way? Are mistakes corrected right at the moment of copying, when *both* strands — master and copy — are available for inspection and comparison?

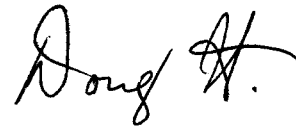
Another question was prompted by something you wrote on page 127: “Binding [of an antigen to an immune receptor] also results in secretion of antibodies with the same variable-region binding-site [the hyphens are mine] and therefore the same antigen specificity as the receptor.” I wrote in the margin that this seems to me to imply an “illegal” information flow (contrary to the Central Dogma, that is) — namely, *from protein* (the immune receptor) *to DNA* (that of the B cell) and then back out to protein (the antibody). This is probably an incorrect image that I have, but I wonder what really goes on. Can you explain this to readers in a bit more detail?

I have one final question, prompted by my interest in so-called “genetic algorithms” (computer models of intelligence inspired by evolution): Is it meaningful to say which is more important as a driving force in evolution — mutation or recombination? My colleague and friend John Holland, at the University of Michigan, maintains that mutation is just a small force in evolution, and that what

really drives it forward is recombination. He has mathematical arguments that in some sense prove this claim, but I wonder if it is a generally accepted notion, or if biologists would dispute it.

Well, finally I have come to the end of this letter. I hope you find my comments helpful and not depressing. Please note that I didn't suggest any global changes — almost everything is pretty local, and therefore not all that hard. I really am looking forward to the final version of the book — it ought to be great! And I'm sure that on my next read-through of it, I will absorb considerably more. It's been fun.

Sincerely,

A handwritten signature in black ink, appearing to read "Doug H." with a stylized flourish at the end.

P.S. — In case you have questions about suggestions I have made, feel free to contact me by letter or by phone (I'm usually at home: (812) 333-4334). I am keeping a photocopy of the whole thing.